PB86-204658

Determination of Chemical Classes from Mass Spectra of Toxic Organic Compounds by SIMCA Pattern Recognition and Information Theory

(U.S.) Environmental Monitoring Systems Lab.
Research Triangle Park, NC

1986

U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO.	2.	3. RE
EPA/600/J-86/064		PB86-204658

4. TITLE AND SUBTITLE	5. REPORT DATE
DETERMINATION OF CHEMICAL CLASSES FROM MASS SPECTRA OF TOXIC ORGANIC COMPOUNDS BY SIMCA PATTERN RECOGNITION AND INFORMATION THEORY (Journal Version)	1986
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S)	8. PERFORMING ORGANIZATION REPORT NO.
D. R. Scott	
9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT NO.
same as #12	L104/F/01/01
	11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS	13. TYPE OF REPORT AND PERIOD COVERED
Environmental Monitoring Systems Laboratory Office of Research and Development U.S. Environmental Protection Agency Research Triangle Park, NC 27711	14. SPONSORING AGENCY CODE EPA/600/08

15. SUPPLEMENTARY NOTES

Analytical Chemistry, 58(4):881-890, April 1986

16. ABSTRACT

The low resolution mass spectra of a set of 78 toxic volatile organic compounds were examined for information concerning chemical classes. These compounds were predominately chloro- and/or bromoaromatics, -alkanes, or -alkenes, which are routinely sought at trace levels in ambient air samples using gas chromatography-mass spectrometry. The set of mass spectra of the pure compounds contained 151 different masses in the range of 35 to 256. The Shannon information content for each mass channel was calculated for the binary encoded and the full intensity spectra, using 1% of the base peak as the threshold level. The 17 masses with the highest Shannon information content were retained as a basis set (the compressed set) for SIMCA pattern recognition using principal component analysis. Examination of the inherent class structure of the compressed set of mass spectra showed a separation of the data into two major classes: aromatics and alkaenes (alkanes and alkenes). The aromatic compounds could be divided into two subclasses: chloro- and nonchloroaromatics. The alkaenes also could be divided into two subclasses: bromo and chloro, with compounds containing both substituents included in the bromo subclass. Except for the total alkaenes class model, the models consisted of one principal component with five masses per component. The total alkaenes model consisted of two principal components with twelve masses. Classification accuracy was 96% for the two major classes and 82% for the four subclasses.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS *(This Report)*	21. NO. OF PAGES
RELEASE TO PUBLIC	UNCLASSIFIED	11
	20. SECURITY CLASS *(This page)*	22. PRICE
	UNCLASSIFIED	

EPA/600/J-86/064
JOURNAL ARTICLE

Published in Analytical Chemistry, April 1986, pp. 881-890, by the American Chemical Society

Determination of Chemical Classes from Mass Spectra of Toxic Organic Compounds by SIMCA Pattern Recognition and Information Theory

Donald R. Scott

Environmental Monitoring Systems Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711

The low-resolution mass spectra of a set of 78 toxic volatile organic compounds were examined for information concerning chemical classes. The Shannon information content for each mass channel was calculated for the binary encoded and the full intensity spectra, using 1% of the base peak as the threshold level. The 17 masses with the highest binary information content were retained as a compressed basis set for SIMCA pattern recognition. The inherent class structure of the data showed two major classes, aromatics and alkaenes (alkanes and alkenes), and four subclasses, chloro- and nonchloroaromatics and bromo- and chloroalkaenes. Except for the total alkaenes class model, the models consisted of one principal component with five masses per component. The total alkaenes model consisted of two principal components with 12 masses. Classification accuracy was 96% for the two major classes and 82% for the four subclasses.

In view of the current data deluge from modern analytical instrumentation, more information is frequently available than is necessary to solve a given problem. The practicing chemometrician should attempt to maximize the use of available analytical information for the solution to a given problem while minimizing the cost of obtaining and processing the data. In the present study the primary goal was to obtain as much information concerning chemical class identification as possible from pattern recognition studies of low-resolution mass spectra of mixtures of trace organic compounds in ambient air. In certain survey studies, e.g., for potential health hazards, or for preliminary screening of mass spectral data files this type of identification is very useful. The data files that would be used in this type of analysis are those available from routine gas chromatography–mass spectrometric analysis of ambient air samples. A secondary goal was to develop procedures that could be performed in the laboratory by analytical chemists using personal computers or small laboratory minicomputers. This requires a compression of the data files to a small number of relevant mass spectral peaks.

The application of pattern recognition to mass spectra has been reviewed recently by Martinsen (*1*). Some of the previous chemometric investigations of mass spectra include factor analysis studies (*2–7*), primarily to determine the number of components in a mixture, cluster analysis and K nearest neighbor studies (*8–12*), discriminant analysis studies (*13, 14*), and SIMCA principal component studies (*12, 15*). The application of information theory to mass spectral data has been reported by Wangen et al. (*16*), by van Marlen, Dijkstra, and van't Klooster (*17–19*), and by van Marlen and van den Hende (*20*). Related studies of the application of information theory to infrared spectral data have been reported by Dupuis, Dijkstra, and van der Maas (*21, 22*) and by Bink and van't Klooster (*23*). The use of information theory in selecting spectral features for retrieval of infrared reference spectra also has been described by Dupuis et al. (*24*) and by Heite et al. (*25*).

Only a few of these chemometric studies were directed toward determination of chemical classes. Rozett and Petersen (*2, 4*) used factor analysis to study the mass spectra of 22 alkylbenzene isomers and to determine the class structure in these compounds. Justice and Isenhour (*3*) used factor analysis to determine the relationship between the phenyl, carboxyl, ether, hydroxyl, nitrogen, amine, and saturated hydrocarbon functional groups and the mass spectra of 453

compounds. Heller, Chang, and Chu (*9*) used cluster analysis to determine the most significant masses for identifying alkylthiol esters in a set of 323 sulfur-containing compounds. Ziemer et al. (*10*) used K nearest neighbor analysis to develop a classification method for amino acids from the mass spectra of 86 dipeptides. Lowry et al. (*11*) used K nearest neighbor methods to assign 20 substructures to a set of 500 mass spectra. Lam et al. (*14*) applied simplex optimization to obtain linear discriminant functions for the phenyl, carbonyl, ether, alcohol, phenol, acid, thiol/thio ether, ester, amine, amide, and nitrile functional groups with a set of mass spectra for 1900 compounds. Wold and Christie (*15*) applied SIMCA pattern recognition to the autocorrelation transformed mass spectra of 21 straight chain and cyclic hydrocarbons. They were able to differentiate between the alkylpentanes, alkylcyclopentanes, and alkylcyclohexanes. Dromey (*26*) has developed a simple series index for classifying mass spectra into 22 compound classes based on an intensity weighted measure of the displacement of the fragment ions from an alkene reference spectrum. In only a few of these studies (*2*, *4*, *15*) was the inherent class structure of the data allowed to define the classes obtained, rather than using preconceived ideas of the classes based on chemical knowledge or intuition.

The present study is concerned with the development of methods for efficient extraction of information regarding chemical class identification from low-resolution mass spectra obtained during routine gas chromatographic–mass spectrometric analysis of volatile trace organic compounds in ambient air samples. The set of 78 toxic compounds investigated contained primarily aromatic compounds, haloalkanes, and haloalkenes with four ethers and epoxides. Approximately 80% of these compounds contain chloro and/or bromo groups. All alkanes and alkenes contained at least one halogen. The methods used were SIMCA pattern recognition (disjoint principal component analysis) with the use of Shannon information content for feature selection from the binary encoded mass spectral data. The data were compressed from an original set of 151 masses to a set of 17 most informative masses. The analysis was performed on a small commercially available 64k CPU microcomputer. It will be shown that this procedure results in the determination of two major classes for the set of compounds with four to five subclasses determined by the number and type of halogens present in the compounds. The use of the binary encoded representation of the mass spectral data vs. full intensity data will be discussed.

THEORETICAL BACKGROUND

Pattern recognition can be construed to mean the application of several different techniques to chemical data (*27*). In this study it is used in the sense of classification of objects into sets based upon some unknown similarity in properties which is inherently present in the data under examination. Classification can be obtained at different levels depending upon the requirements of the analysis (*28*). At the lowest level the objective is to assign an object to one of a set of predefined classes. At the next level an object is to be assigned to one of the predefined classes with the possibility that it may belong to none of the classes. At the highest level an object is classified, and some quantitative information relative to the variables under examination is obtained. In this study pattern recognition will be used at the first level with the 78 toxic compounds as the objects and their mass spectra as the data variables of interest.

For analysis, it is necessary to arrange the mass spectral data into a matrix consisting of n objects (the 78 compounds) arranged in rows with p columns of variables (the 151 peak intensities). The objects are designated with a subscript i, and the variables are designated with a k. An element in the matrix, x_{ik}, represents the value of variable k for object i. Each object in the matrix can be considered to represent a single point in a p-fold hyperspace (measurement space) defined by the row vector of p variables considered. Each of the variables in the row vector represents the value of the coordinate of the object point along the kth axis in this measurement space. If the objects are similar with regard to the variables used, then the points in measurement space should be close together and form a cluster or class.

One of the important functions of principal component analysis is the reduction of dimensionality (compression of variables) to the minimum required for the solution to a given problem. It may also allow an overview or graphical representation of the data set in two-dimensional plots. This allows the user to "see" the relationships between the objects in the data set. This process is accomplished by fitting two or more principal components to the data. The first component is oriented along the axis of greatest variance of the variables in the data matrix about their means. The second principal component is independent of (orthogonal to) the first and is the vector along the axis of next greatest variance in the data. Succeeding principal components can be calculated which will be orthogonal to the preceding ones and which may explain some of the remaining variance. The principal components are linear combinations of the original variables which are fitted in the least squares sense through the points in measurement space. These new variables usually result in a reduction of variables from the original set and often can be correlated with physical or chemical factors. The coefficients of the original variables in the principal components, the loadings, provide information regarding important and redundant variables for the analyzed data.

Pattern recognition is usually carried out in stages involving training and test sets of data. The training or calibration sets of objects are used to define the classes of interest via mathematical models derived from the variables relevant to the classification problem. The test set can be considered as a measure of quality control to verify that the classification models derived from the training set are indeed working correctly.

SIMCA Pattern Recognition. The SIMCA (soft independent modeling of class analogy) pattern recognition techniques were developed by Wold and co-workers and have been described in the literature (*29*, *30*). The statistical pattern recognition techniques are based on disjoint principal component models for classification of objects and canonical partial least squares procedures for establishing quantitative relationships among variables. A version of these procedures, SIMCA 3B, is available which will run on a microcomputer. The computer programs are user interactive and graphically oriented.

The SIMCA class models are bilinear projection models obtained by decomposing the class data matrix $|\mathbf{X}|$ into a score matrix $|\mathbf{T}|$ ($n \times F$), a loading matrix $|\mathbf{P}|$ ($F \times p$), and a residual matrix $|\mathbf{E}|$ (*30*)

$$|\mathbf{X}| = 1 \cdot \bar{\mathbf{x}} + |\mathbf{T}||\mathbf{P}| + |\mathbf{E}| \qquad (1)$$

The row vector $\bar{\mathbf{x}}$ is composed of all the averages of the variables in the class data matrix. The $n \times F$ score matrix $|\mathbf{T}|$ describes the projection of the n object points down on the F dimensional hyperplane defined by the $F \times p$ loading matrix $|\mathbf{P}|$. The residual matrix $|\mathbf{E}|$ contains that part of the data matrix due to measurement and modeling errors. If the residuals in $|\mathbf{E}|$ are small compared with the variation in $|\mathbf{X}|$, then the model is a good representation of $|\mathbf{X}|$. If the dimension, F, of the hyperplane is smaller than p, the original number of variables, then a reduction of dimensionality (number of variables) has been achieved. When F is two or three, the columns in the score matrix $|\mathbf{T}|$ can be plotted

against each other to get two-dimensional pictures of the objects in hyperspace (measurement space). The modeling power, i.e., the reduction of variance, for a given variable can be used as a measure of its relevance to the class model. Judicious use of the variable loadings and modeling powers obtained in preliminary analyses of the data allows one to polish the model and remove variables that are not significant to the class.

The number of principal components determined and retained for a particular class model is an important consideration. Within the SIMCA procedures the process of cross validation (*31*) is used to determine the number of statistically significant principal components for a given class. In this process subsets of the training set of objects are used to fit a class model with one principal component. The objects not used in determining this class model are then fitted to the model, and the sum of squared residuals for the withheld objects is calculated. This is repeated until all of the objects in the training set have been withheld from the model fitting. The overall sum of squared residuals is then calculated for all the withheld objects for this particular model. The entire process is repeated with a class model containing an additional principal component. Addition of principal components to the model is continued until comparison of the sums of squared residuals for the previous and present model show no improvement. Generally, if the number of principal components for a given class model is much smaller than the number of either objects or variables for the class; then the model will be statistically stable.

Once the class models have been determined, objects are classified by fitting their data to the various class models. A standard deviation for each model is calculated from the residuals. This represents a class tolerance level around the principal component model in measurement space. The standard deviations for the objects are calculated from the residuals, and the objects are classified based upon their distances from the class models.

Shannon Information Theory. The Shannon information content of a message is related to the reduction in uncertainty gained by the receiver of the message. The information content will depend on the probability of occurrence of the symbols used, correlation between the symbols, and encoding and decoding errors. In low-resolution mass spectrometry the message is equated to the mass spectrum itself. The symbols are the intensities at a given unit mass channel. Neglecting errors, the information content per mass channel, $I(j)$, is given by

$$I(j) = -\sum_{i=1}^{m} p_j(i) \log_2 p_j(i) \qquad (2)$$

where m is the number of discrete intensity values available for mass channel j. The probabilities of occurrence of a mass intensity, $p_j(i)$, are calculated over the entire set of reference spectra. The total information content of a given mass spectrum is the sum of the information contents of the individual channels if correlation between masses is neglected. Since correlation between masses is well-known in mass spectra, this information content should be regarded as an upper limit to the true information content.

For binary encoded spectra there are only two possibilities for the intensity of a mass peak, 1 or 0, corresponding to the presence or absence of the peak above a threshold level. Therefore the information content for mass channel j can be calculated from

$$I(j) = -p_j \log_2 p_j - (1 - p_j) \log_2 (1 - p_j) \qquad (3)$$

Here p_j is the probability of a peak occurring at the mass channel j calculated over the set of mass spectra of the 78 compounds. The maximum information content per channel for the binary case is one bit and occurs at a probability of 0.5.

Hamming Spaces. The data vector for an object can be considered to represent a point in the p-dimensional measurement space. With the full intensity data the limits of this space are represented by a p-dimensional hypercube with an edge length equal to the intensity of the most intense peak, the base peak. This length can be set equal to one. The interior of this hypercube will be occupied by the object points in the data set as their coordinates dictate. Distances between points can be calculated from the Euclidean or other formulas.

In the case of binary encoded data the data vector for an object represents a point in a p-dimensional Hamming space, which again can be visualized as a p-dimensional hypercube with unit edge. However, since only values of 0 or 1 are assigned to peak intensities, all of the object points lie only on the corners of the hypercube. The interior of the space is not occupied. The distance between two points in the Hamming space is the minimum distance along the edges of the hypercube between the two vertices where the points lie. The Hamming distance is equal to the number of mismatches between the binary encoded data vectors (mass spectra) being compared (*32*). Therefore the Hamming distance is related to the logical exclusive OR operator (XOR). The Hamming distance is the square of the corresponding Euclidean distance between two points in this space.

EXPERIMENTAL SECTION

Data Set. The low-resolution mass spectra of the 78 compounds were obtained from the EPA-NIH Mass Spectral Library on an INCOS data system. A typical spectrum contained approximately 16 peaks. The range of mass/charge ratios was from 35 to 256 with 151 different peaks occurring in the total set. For the full intensity data the intensities were scaled to give a maximum of one for the base peak. The data were also binary encoded by assigning an intensity of one to any peak over the threshold level, which was 1% of the base peak intensity. A list of individual compounds in the set is given in Table I.

Hardware and Software. In this study an Osborne 1 microcomputer (Z80A) with a CP/M operating system and 64k memory was used. This amount of memory is sufficient to handle a data matrix of size 50 objects by 50 variables. The program occupied 220k space on double density floppy disks. The SIMCA 3B software package was obtained from Principal Data Components, Columbia, MO. It included modules to define a data file; to scale, weight, and transform data; to edit, merge, or split the data file; to list the file; to input the data, define classes, and perform K nearest neighbor analysis; to plot the data; to perform principal component analysis for classes; to test the fit of data to defined classes; and to predict values of dependent variables from relationships with independent variables with partial least squares.

Data Analysis. The information content of the 151 different nonzero intensity mass channels was calculated from the binary encoded mass spectra of the set of 78 compounds using the formulas given above. The full intensity mass spectra of the data set were encoded into ten discrete levels above the threshold level, and the information content for the mass channels also was calculated using this distribution. No correction was made to account for encoding errors or for correlation between mass channels. Therefore the calculated information contents should represent upper limits to the true values.

The class modeling and object classification were performed with the SIMCA 3B software. The data were preprocessed by binary encoding of only those 17 mass channels selected after calculation of the Shannon information content. All data were class scaled. The training sets for each of the classes were selected after investigating the inherent structure of the data set. All objects in a given class were used for the training sets, except for bromobenzene, which was only used in the total aromatic training set. The numbers of compounds used in the training sets for the classes are shown in Table II. Cross validation was used to determine the number of statistically significant components for

Table I. Compounds Included in This Study

1	*p*-xylene	40	bromoform
2	1,3,5-trimethylbenzene	41	1,2-dichloroethane
3	isopropylbenzene	42	1,1,1-trichloroethane
4	*n*-butylbenzene	43	1,1,2-trichloroethane
5	1-methyl-4-isopropylbenzene	44	1,1,2,2-tetrachloroethane
6	*o*-dichlorobenzene	45	pentachloroethane
7	*p*-dichlorobenzene	46	1,1-dichloroethene
8	1-chloro-2-methylbenzene	47	trichloroethene
9	1-chloro-4-methylbenzene	48	tetrachloroethene
10	*p*-chlorostyrene	49	bromoethane
11	1,1-dichloroethane	50	1,2-dibromoethane
12	1,1,1,2-tetrachloroethane	51	1-chloropropane
13	1,2,3-trichloropropane	52	2-chloropropane
14	3-chloropropene	53	1,2-dichloropropane
15	2-chlorobutane	54	1,3-dichloropropane
16	1,3-dichlorobutane	55	1-bromo-3-chloropropane
17	1,4-dichlorobutane	56	1,2-dibromopropane
18	1,4-dichloro-2-butene (cis)	57	2,3-dichlorobutane
19	3,4-dichlorobutene	58	tetrahydrofuran
20	1,4-dioxane	59	benzaldehyde
21	1-chloro-2,3-epoxypropane	60	1-bromo-1-chloroethane
22	2-chloroethoxyethene	61	2,2-dibromopropane
23	acetophenone	62	2-bromopropene
24	benzonitrile	63	2-bromopropane
25	benzene	64	3-bromopropene
26	toluene	65	1-bromopropane
27	*o*-xylene	66	1-chlorobutane
28	*m*-xylene	67	1-bromo-2-chloroethane
29	ethylbenzene	68	bromodichloromethane
30	styrene	69	1-bromobutane
31	chlorobenzene	70	2,2-dichlorobutane
32	bromobenzene	71	dibromochloromethane
33	*m*-dichlorobenzene	72	1,1,2-trichloropropane
34	1-chloro-3-methylbenzene	73	1,3-dibromopropane
35	chloroform	74	1,1,1,2-tetrachloropropane
36	carbon tetrachloride	75	1,2,2,3-tetrachloropropane
37	bromochloromethane	76	1,3-dibromobutane
38	bromotrichloromethane	77	1,1,2,3-tetrachloropropane
39	dibromomethane	78	1,4-dibromobutane

Table II. Number of Compounds in Chemical Classes and Training Sets[a]

substituent	aromatics	alkaenes[b]	others[c]
no Cl or Br	14 (14)	0	2
Br	1	14 (21)[e]	0
Cl	8 (8)	30 (30)	2
Br and Cl[d]	0	7[e]	0
total	23 (23)	51 (51)	4

[a] Numbers in parentheses are numbers of compounds used in training sets. [b] Alkaenes includes alkanes and alkenes. [c] Others includes ethers and epoxides. [d] Substituted with both Cl and Br. [e] The seven Br- and Cl-alkaenes were used in the training set for the Br-alkaene class.

each class model. Variables with modeling powers less than 0.18 for the principal component models were deleted in the initial stages of the refinement of the class models.

RESULTS AND DISCUSSION

Mass Information Content and Feature Selection. The information content for the 151 mass channels from the binary encoded spectra are listed in Table III along with those reported for a set of 9600 binary encoded spectra (*18*). The information content per channel for the full intensity mass spectra also is given in Table III. The binary encoded spectra of the 78 compounds yielded information contents per channel of 0.10 to 1.00 bit with the higher information at mass channels below 107. Of the channels from 107 and below, 70% had 0.5 bit or greater information. Of the channels greater than 107, only 7.4% had 0.5 bit or greater information content. This concentration of higher information content in lower mass channels is related to the more frequent occurrence, and higher probability, of low masses. This has been noted in previous studies (*18*).

The information content for the full intensity spectra of the 78 compounds ranged from 0.10 to 2.29 bit with no channels above 107 exceeding 1.02 bit. In general the information contents of the binary and full intensity mass spectra show a linear correlation up to a binary information content of ca. 0.9 bit. Even in the case of the higher information channels there is good qualitative agreement as shown in Table IV where the 18 mass channels with the highest binary spectral information content are compared with those from the full intensity spectra. Of these 18 mass channels 16 are also found in the 18 highest information content channels for the full intensity spectra. Comparison of the information content of these 18 mass channels for the binary spectra of the 78 compounds with those found for the set of 9600 compounds also shows a good correlation. The 18 highest information mass channels for the set of 78 compounds yielded 0.80–1.00 bit with a median of 0.92 bit, while the same mass channels with the set of 9600 compounds gave 0.62 to 1.00 bit with a median of 0.91 bit. Thus it is clear that this set of 18 most "informative" mass channels contains very much information not only with regard to the set of 78 compounds but also with regard to the set of 9600 compounds.

It is well-known that the information in a complete mass spectrum is highly redundant and that compressed binary encoded spectra retain a large portion of the information present in the complete spectrum (*4, 13, 16, 18, 33*). As a result it is not necessary to use the full spectrum in pattern recognition studies, and in fact the use of the full spectrum may obscure class definition inherent in the data structure. To solve the present pattern recognition problem on a microcomputer, the number of mass channels considered was reduced from 151 to the 17 channels listed in Table IV. These channels have the highest information content, i.e., greater than 0.80 bit, for the binary encoded spectra with the exception of channel 39. This latter mass channel was not used since it is one of the most frequently occurring peaks in mass spectrometry (*34*). This set of binary encoded, 17 mass channel spectra, which will be designated hereafter as the compressed spectra, was used as the trial set of variables in the pattern recognition classification of the 78 compounds. The use of the compressed set of binary spectra causes nine pairs and one trio of compounds to have identical spectral representations.

Inherent Structure of the Data Set. As pointed out by Wold and Christie (*15*) the inherent structure of the data should determine the nature of the classes and not some preconceived scheme based on chemical training or intuition. The number and types of classes present in a data set will depend upon the particular variables selected for the study. For a preliminary overview of the data a two-dimensional principal component plot of the 17 mass, full intensity data was constructed. Only 49 of the 78 compounds were considered in this plot due to limitations of the SIMCA program used. However, the compounds selected included representatives of all apparent chemical classes in the set. There was no class separation of any type visually apparent in the resulting plot. This result was not unexpected since it has been pointed out in previous studies (*12, 15*) that untransformed mass spectral data should not be used for pattern recognition studies.

Suspecting that the use of the full intensity data might be obscuring the underlying basic structure of the data, we decided to examine the binary encoded, 17 mass data for the same 49 compounds. A two-dimensional principal component plot of the results is shown in Figure 1. It is apparent from this plot that there is some basic separation of the compounds

Table III. Information Content of Mass Channels for Data Set

mass	bits			mass	bits			mass	bits			mass	bits		
	I(B)[a]	I(9600)[b]	I(D)[c]		I(B)	I(9600)	I(D)		I(B)[a]	I(9600)[b]	I(D)[c]		I(B)	I(9600)	I(D)
35	0.55	0.36	0.72	86	0.17	0.73	0.17	135	0.48	0.60	0.68	169	0.10	0.47	0.10
36	0.52	0.46	0.57	87	0.39	0.76	0.46	136	0.34	0.49	0.41	170	0.17	0.32	0.20
37	0.76	0.77	0.83	88	0.34	0.56	0.47	137	0.29	0.55	0.37	171	0.23	0.39	0.27
38	0.78	0.95	1.00	89	0.52	0.81	0.68	138	0.29	0.47	0.37	172	0.10	0.33	0.10
39	0.98	0.83	2.29	90	0.52	0.58	0.66	139	0.10	0.64	0.10	173	0.23	0.37	0.27
40	0.65	1.00	0.71	91	0.92	0.96	1.47	140	0.10	0.51	0.10	174	0.10	0.33	0.10
41	0.92	0.90	1.92	92	0.73	0.81	0.98	141	0.10	0.63	0.10	175	0.23	0.39	0.27
42	0.73	0.99	0.91	93	0.87	0.83	1.22	142	0.17	0.47	0.20	176	0.10	0.37	0.10
43	0.62	0.95	0.98	94	0.48	0.71	0.59	144	0.29	0.46	0.33	186	0.10	0.29	0.10
44	0.39	0.99	0.44	95	0.90	0.83	1.39	145	0.23	0.55	0.27	188	0.10	0.29	0.10
45	0.34	0.94	0.39	96	0.52	0.74	0.67	146	0.48	0.42	0.64	190	0.10	0.32	0.10
46	0.10	0.54	0.10	97	0.71	0.85	1.17	147	0.34	0.54	0.39	200	0.23	0.30	0.27
47	0.62	0.55	0.84	98	0.55	0.77	0.72	148	0.29	0.44	0.33	201	0.10	0.31	0.10
48	0.34	0.34	0.39	99	0.68	0.75	0.97	149	0.23	0.50	0.23	202	0.23	0.34	0.27
49	0.98	0.62	1.68	100	0.23	0.55	0.23	150	0.23	0.46	0.23	203	0.10	0.34	0.10
50	0.84	0.98	1.18	101	0.34	0.69	0.41	156	0.17	0.36	0.20	204	0.17	0.29	0.20
51	0.99	1.00	1.78	102	0.34	0.69	0.39	157	0.10	0.44	0.10	206	0.10	0.26	0.10
52	0.59	0.92	0.68	103	0.59	0.80	0.77	158	0.29	0.37	0.37	208	0.10	0.26	0.10
53	0.65	1.00	0.84	104	0.55	0.72	0.61	159	0.17	0.43	0.17	210	0.10	0.22	0.10
54	0.29	0.91	0.33	105	0.78	0.86	1.12	160	0.29	0.37	0.33	212	0.10	0.22	0.10
55	0.65	0.99	0.96	106	0.65	0.70	0.99	161	0.10	0.45	0.10	214	0.17	0.22	0.17
56	0.52	0.99	0.66	107	0.82	0.74	1.08	162	0.17	0.38	0.17	216	0.17	0.28	0.20
57	0.44	1.00	0.57	108	0.34	0.62	0.39	163	0.10	0.48	0.10	218	0.17	0.22	0.17
58	0.34	0.94	0.39	109	0.71	0.74	1.00	164	0.10	0.38	0.10	250	0.10	0.16	0.10
59	0.34	0.81	0.34	110	0.44	0.67	0.56	165	0.23	0.51	0.30	252	0.10	0.20	0.10
60	0.55	0.66	0.72	111	0.48	0.74	0.66	166	0.17	0.41	0.20	254	0.10	0.19	0.10
61	0.82	0.75	1.36	112	0.52	0.68	0.66	167	0.23	0.45	0.27	256	0.10	0.20	0.10
62	0.95	0.81	1.61	113	0.39	0.67	0.50	168	0.17	0.41	0.20				
63	1.00	0.97	1.95	114	0.17	0.47	0.20								
64	0.68	0.79	0.94	115	0.23	0.82	0.23								
65	0.92	0.98	1.35	117	0.44	0.75	0.60								
69	0.10	1.00	0.10	118	0.10	0.63	0.10								
70	0.10	0.94	0.10	119	0.52	0.76	0.73								
71	0.10	0.95	0.10	120	0.55	0.60	0.77								
72	0.10	0.76	0.10	121	0.68	0.68	1.02								
73	0.55	0.82	0.64	122	0.44	0.56	0.57								
74	0.59	0.86	0.73	123	0.44	0.63	0.56								
75	0.89	0.89	1.42	124	0.39	0.54	0.50								
76	0.62	0.83	0.89	125	0.23	0.63	0.23								
77	0.94	1.00	1.75	126	0.44	0.62	0.56								
78	0.84	0.93	1.31	127	0.39	0.73	0.44								
79	0.96	0.97	1.27	128	0.39	0.67	0.50								
80	0.48	0.74	0.48	129	0.39	0.67	0.53								
81	0.80	0.91	1.02	130	0.29	0.54	0.37								
82	0.55	0.88	0.67	131	0.59	0.65	0.85								
83	0.59	0.92	0.89	132	0.44	0.53	0.57								
84	0.44	0.85	0.54	133	0.39	0.63	0.54								
85	0.62	0.90	0.86	134	0.44	0.53	0.60								

[a] Information content of binary encoded spectra of 78 compounds. [b] Information content of binary encoded spectra of 9600 compounds (ref 18). [c] Information content of full intensity spectra of 78 compounds.

into at least three classes and probably four. In the lower right corner is a bromo-substituted group of alkenes and alkanes (hereafter collectively designated as alkaenes) including bromochloro-substituted alkaenes. In the lower left corner is a group of nonhalogenated aromatic compounds. At the top center is a group of chloroalkaenes. Lying between and partially overlapping the latter two groups is a group of chloroaromatics. An alternant three category classification scheme would consist of nonhalogenated aromatics, bromoalkaenes, and chlorinated alkaenes and aromatics.

Modeling of the Major Mass Spectral Classes. It is clear that the compressed set of data contains enough information for useful classification of the 78 compounds and that the classification scheme should concentrate on chloro- or bromo-substituted alkaenes and chloro- or nonhalogenated aromatics. A classification matrix based on this general scheme and including the numbers of compounds in each category and training set is given in Table II. It is somewhat a matter of choice and a test of the modeling ability of the data as to the number of classes used. We have chosen to use two general classes: aromatics with 23 compounds in the training set and alkaenes with 51 compounds in the training set. For further detailed classification four subclasses were used: nonhalogenated aromatics with 14 training set compounds; chloroaromatics with 8 training set objects; chloroalkaenes with 30 training set compounds; and bromoalkaenes with 21 training set compounds. The seven alkaenes with both chloro- and bromo-substituents were found during initial calculations to fit with the bromoalkaenes. The additional four compounds, the three ethers and one epoxide, were withheld from the training sets and used as test objects.

With the use of the variable modeling power and the compressed set of 17 mass spectra, it was found that only five masses were required for each refined class model except for the total alkaenes class. In the total alkaenes case 12 masses were required. The use of cross validation resulted in the refined models having only one principal component per model except for the total alkaenes model which had two. The model

Table IV. Most Informative Masses for Data Set

mass	bits I(B)[a]	bits I(9600)[b]	bits I(D)[c]	mass	bits I(B)	bits I(9600)	bits I(D)
39[d]	0.98	0.83	2.29	75	0.89	0.89	1.42
41	0.92	0.90	1.92	77	0.94	1.00	1.75
49	0.98	0.62	1.68	78	0.84	0.93	1.31
50	0.84	0.98	1.18	79	0.96	0.97	1.27
51	0.99	1.00	1.78	81	0.80	0.91	1.02
61	0.82	0.75	1.36	91	0.92	0.96	1.47
62	0.95	0.81	1.61	93	0.87	0.83	1.22
63	1.00	0.97	1.95	95	0.90	0.83	1.39
65	0.92	0.98	1.35	107	0.82	0.74	1.08

[a] Information content of binary encoded spectra of 78 compounds. [b] Information content of binary encoded spectra for 9600 compounds. [c] Information content of full intensity spectra of 78 compounds. [d] Not included in set of masses selected for compressed set.

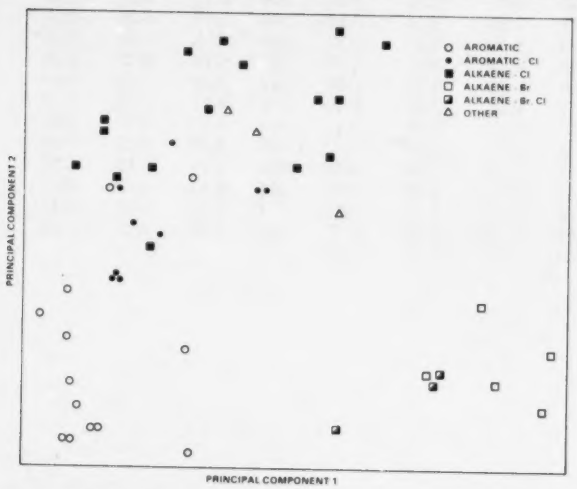

Figure 1. Principal component plot of binary encoded 17 mass spectra of 49 compounds.

parameters including loadings, variable modeling power, unexplained variance, and residual standard deviation are listed in Table V. The unexplained variance ranged from a low of 19% for the chloroaromatic model to a high of 47% for the chloroalkaene model. The residual standard deviations, which are a measure of the spatial extent of the models, ranged from 0.49 for the chloroaromatic model to 0.77 for the chloroalkaene model.

Classification of Compressed Spectra. After the class models were determined, the entire set of 78 compressed spectra was run through all of the six models to determine classification accuracy. Compounds were considered as members of the class to which they were closest. The results are given in Table VI as training set accuracy, overall accuracy, and percent misidentified as a class member for the six classes. Classification results for the test set compounds, which do not properly belong to any of the classes, are also given. In the case of the two major classes, aromatics and alkaenes, the training set accuracy was 91 and 98%, respectively, with an overall accuracy of 96%. Only 1 and 3% of the 74 compounds were incorrectly classified as members of the two classes. All of the test set compounds were classified as alkaenes, which is correct when interpreted as "not aromatic".

For the four subclasses, nonchloroaromatics, chloroaromatics, bromoalkaenes, and chloroalkaenes, the training set accuracy was 79, 87, 86, and 80%, respectively. The overall accuracy was 82%. The number of compounds misidentified as members of the four classes were 3, 4, 3, and 8%, respectively. In this case bromobenzene was added to the four test set compounds. The two chloro-substituted alkaene type compounds were correctly classified. Bromobenzene was classified as a chloroaromatic which is the best of the four possible classes. The two cyclic alkane ethers were classified as bromoalkaenes, which if regarded as "alkaene, but not chloroalkaene", is correct within the available models.

Detailed Chloroalkaene Models and Classification. It was also possible to construct a class model for chlorinated alkaenes based upon the number of chloro groups present in a compound. A single principal component model using the five masses—63, 65, 75, 77, 91—was found to be statistically valid for the class of mono- and dichloroalkaenes. This model did not include compounds substituted with both chloro and bromo groups. This model gave an unexplained variance of 42% for the 16 training set compounds with a residual standard deviation of 0.73. The loadings of the principal component were ca. 0.4–0.5 with mass 63 and 65 having negative loadings. The training set accuracy was only 62%, but the overall accuracy for 78 compounds was 91%. Only two compounds, both monochloromonobromoalkaenes, were misidentified as members of this class. These results were obtained by using a maximum class distance of 0.76, a little larger than the class residual standard deviation.

Attempts to model the 13 remaining chloroalkaenes containing three or more chloro substituents per compound as a class did not result in a statistically valid model.

Interpretation of Particular Masses in Models. The specific masses found to be important in these models may be specific to the set of compounds investigated and are obviously restricted by the 17 masses chosen initially. Generally the models were simple sums and differences of masses as dictated by the principal component models. Sixteen of the 17 most informative masses were found in the refined class models with 12 being used in the all alkaenes model alone. Mass channel 78 was used in none of the classes. Mass channels 63 and 65 were the most prevalent, appearing in four and five, respectively, of the six major classes. Channels 79, 81, 93, 95, and 107 were only found in one class model, primarily the all alkaenes model. The main difference between the all aromatic and all alkaene models was the appearance of masses 65 and 91 in the former but not in the latter model. The principal difference between the nonchloroaromatic and chloroaromatic models was the exclusive appearance of masses 49 and 75 in the former and 62 and 63 in the latter. The bromo- and chloroalkaene models were more varied with masses 49, 61, 62, and 63 exclusively in the former and 41, 51, 77, and 95 in the latter.

It would appear that the masses in the models should have high information content for the particular classes being modeled. To test this hypothesis the information contents of the 17 mass channels were calculated using binary encoded spectra for each of the six major classes found in the data set. The results including the total information contents for the masses used for the particular models are shown in Table VII. For every class model except the bromoalkaene case, the masses used in the model had at least 0.75 bit of information relative to that class. However, some high information masses did not appear in class models as might be expected. In the bromoalkaene model the results were odd since the masses used had only 0.28–0.79 bit of information. Examination of the spectra of the individual compounds in the bromoalkaene training set showed that the model was constructed on the absence, rather than the occurrence, of the masses found in the model. This led to the relatively low information content of the masses in the model relative to the class. The total information contents per model were ca. 4.4–4.8 bits except for the all-alkaenes model with ca. 11 bit and the bromo-

Table V. Major Class Parameters for Compressed Mass Spectral Data

All Aromatics (N = 23)

Principal Component 1

mass channel	50	63	65	75	91
loading	−0.39	0.34	0.51	−0.45	0.52
model power	0.28	0.19	0.63	0.42	0.65
variance remaining	35%				
residual std dev	0.66				

All Alkaenes (N = 51)

Principal Component 1

mass channel	41	49	51	61	62	63	75	77	79	81	93	107
loading	0.0	0.29	0.25	0.21	0.34	0.29	0.28	0.29	−0.30	−0.37	−0.37	−0.30
model power	0.0	0.24	0.17	0.11	0.35	0.23	0.21	0.23	0.25	0.44	0.46	0.26
variance remaining	59%											
residual std dev	0.80											

Principal Component 2

mass channel	41	49	51	61	62	63	75	77	79	81	93	107
loading	−0.51	−0.04	−0.44	0.43	0.12	0.11	−0.30	−0.41	−0.13	−0.06	−0.12	−0.19
model power	0.28	0.23	0.43	0.33	0.36	0.24	0.32	0.48	0.26	0.44	0.48	0.30
variance remaining	44%											
residual std dev	0.72											

Nonchloroaromatics (N = 14)

Principal Component 1

mass channel	49	50	65	75	91
loading	0.43	0.36	−0.47	0.49	−0.48
model power	0.42	0.25	0.55	0.63	0.58
variance remaining	28%				
residual std dev	0.60				

Chloroaromatics (N = 8)

Principal Component 1

mass channel	50	62	63	65	91
loading	−0.38	0.38	0.49	0.49	0.49
model power	0.33	0.33	0.86	0.86	0.86
variance remaining	19%				
residual std dev	0.49				

Bromoalkaenes (N = 21)

Principal Component 1

mass channel	49	61	62	63	65
loading	0.34	0.49	0.39	0.49	0.49
model power	0.25	0.79	0.36	0.79	0.79
variance remaining	22%				
residual std dev	0.52				

Chloroalkaenes (N = 30)

Principal Component 1

mass channel	41	51	65	77	95
loading	0.52	0.44	0.42	0.40	−0.45
model power	0.47	0.30	0.27	0.23	0.32
variance remaining	47%				
residual std dev	0.77				

alkaene model with a low 2.4 bit. If only four classes were possible for a given compound, then only two bits of information would be required to classify the compound. Even the bromoalkaene model had this amount of information, and the total alkaenes model contained enough information to theoretically distinguish among ca. 2000 categories.

Data Transformation by Binary Encoding. The use of binary encoded mass spectral data had a major effect on the results of this study. It allowed the discovery of the basic classes in the data set, which was not possible with the full intensity data. Physically the use of binary encoded data corresponds to the indication of the presence (value 1) or absence (value 0) of a particular property and, therefore, conveys information at the most fundamental level. Binary data are commonly used in human pattern recognition whether knowingly or not. In the present case it indicates the presence or absence of particular mass channels in the mass spectrum of a compound above a threshold level. The successful use of binary encoded data in pattern recognition studies of this type emphasizes the highly correlated and redundant nature of the complete mass spectrum. The removal of correlations between the intensities and the respective mass positions in the spectrum, even with only 17 masses present, may have been accomplished by the use of the binary data.

Mathematically, binary encoding produces two important results. It reduces the variation of the variable values by assigning one value (1) to all above the threshold level. It also

Table VI. Classification of 78 Compounds into Major Classes

	all aromatics (N = 23)	all alkaenes (N = 51)
% correct (training set)	91	98
% misidentified into class (74 compounds)	1	3
test set		
p-dioxane		X
tetrahydrofuran		X
1-chloro-2,3-epoxypropane		X
2-chloroethoxyethene		X

% Correct Overall for 74 Compounds into Two Classes, 96

	aromatic (N = 14)	chloroaromatic (N = 8)	bromoalkaene[a] (N = 21)	chloroalkaene (N = 30)
% correct (training set)	79	87	86	80
% misidentified into class (73 compounds)	3	4	3	8
test set				
bromobenzene		X		
p-dioxane			X	
tetrahydrofuran			X	
1-chloro-2,3-epoxypropane				X
2-chloroethoxyethene				X

% Correct Overall for 73 Compounds into Four Classes, 82

[a] Bromoalkaenes include those compounds with both bromo and chloro substitution as well as those with only bromo substitution.

Table VII. Information Content of 17 Masses for Classes

mass	information content (bit)[a]					
	all aromatics	all alkaenes	bromoalkaenes	chloroalkaenes	non-chloroaromatics	chloroaromatics
41	0.56	0.98*	1.00	0.95*	0.75	0.00
49	0.67	1.00*	0.79*	0.84	0.75*	0.00
50	0.93*	0.46	0.28	0.57	0.94*	0.95*
51	0.67	0.85*	0.45	0.97*	0.00	1.00
61	0.26	0.94*	0.45*	1.00	0.37	0.00
62	0.67	0.99*	0.28*	0.84	0.59	0.81*
63	0.89*	0.97*	0.45*	0.97	0.37	0.95*
65	1.00*	0.85	0.45*	0.97*	0.99*	0.95*
75	0.99*	0.75*	0.00	0.95	0.94*	0.54
77	0.89	0.79*	0.28	0.95*	0.37	0.81
78	0.99	0.63	0.45	0.72	0.75	0.54
79	0.99	0.94*	0.79	0.35	0.94	0.00
81	0.26	0.94*	0.59	0.00	0.00	0.00
91	0.99*	0.82	0.96	0.65	0.86*	0.95*
93	0.43	0.98*	0.28	0.21	0.59	0.00
95	0.00	1.00	0.86	0.92*	0.00	0.00
107	0.76	0.85*	0.92	0.00	0.94	0.00
class total[b]	4.80	10.98	2.42	4.76	4.48	4.61

[a] The masses with an asterisk were used in the principal component models for the particular classes. [b] This is the sum of the information contents for the masses used in the particular model.

increases the relative magnitude of most peaks above the threshold level. Let us assume that the most intense peak (base peak) in the spectrum has been assigned the intensity value 1. The binary encoding transformation assigns all intensities greater than threshold values, e.g., 0.02, 0.1, 0.5, 0.99, the value 1. This increases the magnitude of all peak intensities except the base peak and any other peaks which are accidentally as intense as the base peak. The result is a smoothing of the intensities to the same value of 1 and an equal weighting, with regard to intensity, of all mass channels above the threshold level. The intensities of the base peaks and those below the threshold level remain unchanged by the transformation.

Geometrically, this binary encoding transformation can be visualized as a shifting of the object points from the interior and edges of the p-dimensional hypercube (with unit edge) in measurement space to the corners of a hypercube in a Hamming measurement space. For the full intensity mass spectral data, which always includes integer variables (0 or 1); the object points would lie on the edges of the p-dimensional hypercube before encoding. For the compressed spectra some objects might lie in the interior of the p-dimensional hypercube since the integer variables might be deleted in the process of compression. After binary encoding, these object points will lie at the corners of the p-dimensional hypercube in Hamming space since only the corners of the hypercube are valid coordinate points, not the edges or interior. The geometric effects of the binary transformation are summarized in Table VIII.

Distances between points in the Hamming measurement space are different from those in the untransformed measurement space. This has an effect on any classification process that is based on distances in the measurement space as measures of similarity. Table VIII shows the changes in separation due to binary encoding. In the table and in the following discussion the distances referred to are the square

Table VIII. Geometric Effect of Binary Encoding Object Vectors

	geometric location on hypercube in	
	measurement space	Hamming space
data vector		
all integers	corners	corners
all nonintegers	interior	corners
integers and nonintegers	edges	corners

	distances between objects[a] in	
	measurement space	Hamming space
mass spectra (P total peaks, N above threshold, B binary mismatches)		
similar spectra ($B = 0$)	$0 \le D < N$	0
nonsimilar spectra ($B \neq 0$)	$0.0001B \le D \le B$	B
after autoscaling (weights W_k)[b]		
similar spectra	$0 \le D < \sum_N W_k^2$	0
nonsimilar spectra	$0.0001\sum_B W_k^2 \le D \le \sum_B W_k^2$	$\sum_B W_k^2$

[a] The distance, D, used is the square of the Euclidean distance to correspond to the Hamming distance. [b] The weights W_k are the autoscaling weights for column k in the data matrix.

of the Euclidian distances. If objects have the same mass spectra except for intensity variations among the N masses above the threshold level, the distance between them, which initially may be as large as N, will decrease to zero upon binary encoding. They will occupy the same corner in the Hamming space. If two objects have B mismatches with regard to numbers of peaks above and below the threshold, their separation, which may be as small as $0.0001B$, will be increased upon binary encoding to a value of B. In the Hamming space their separation will be determined only by the number of total mismatches between above and below threshold peaks.

Therefore, the general effects of binary encoding are (a) to bring into coincidence all objects with similar spectra (differing only in peak intensities above threshold) and (b) to increase the separation between objects with dissimilar spectra. The separation between objects will be determined by the number of mismatches between the spectra (the Hamming distance).

The effect of autoscaling on the data must also be considered. In autoscaling the data, each column k in the data matrix is multiplied by the weighting constant, W_k. In geometric terms this distorts the original unit edged hypercube into a p-dimensional rhombic hyperparallelepiped with edges equal to the respective W_k weighting factors. Before binary encoding, objects are located at the corners, at the edges, and in the interior of this solid figure as was the case before autoscaling. The binary transformation again moves the object points to the corners of the p-dimensional Hamming solid figure. Thus the symmetry of the original hypercube is distorted by the autoscaling process. However objects with similar spectra will still be coincident in this Hamming space, and the separation between objects with dissimilar spectra will still be determined by the number of mismatches between their spectra. Comparison of the bounds on the separation between objects shown in Table VIII shows that after autoscaling, the bounds change by the replacement of N (the number of peaks above threshold) and B (the number of binary mismatches) with the sums of squares of the weights over the N peaks or over the B mismatches, respectively. The importance of the number of mismatches between the two spectra in determining the separation between the two corresponding objects is apparent.

CONCLUSIONS

The use of information theory and SIMCA pattern recognition with the binary encoded mass spectra of the 78 compounds has led to the determination of six classes of compounds in the data set. The four detailed classes are defined as nonchloroaromatic, chloroaromatic, bromoalkaene, and chloroalkaene compounds. A subclass of alkaenes containing one or two chloro substituents was also found. In all models except the all-alkaene model only one principal component was found to be statistically significant. The individual principal components, except for the all alkaene model, were found to contain only five masses of the 17 chosen as a trial set based on high Shannon information content. It should be emphasized that neither the masses selected for the compressed set nor those in the models were chosen using preconceived ideas regarding their class discriminatory powers. The former were selected according to their information content relative to the entire set of mass spectra, and the latter were determined from the modeling powers in the refinement of the class models. The overall accuracy was 96% for alkaene vs. aromatic classification and 82% for classification into the four groups. There also was some information in the data set concerning the number of chloro substituents in the alkaene compounds.

The use of Shannon information content appears to be a compact and powerful way of expressing probability distribution information on spectral data and allowing efficient preprocessing of data for pattern recognition studies. In the present case 151 different mass channels were compressed into 16 which were actually used in the different models. The requirement of high Shannon information content for a particular spectral channel as a condition for importance in SIMCA models is generally supported in this study. However, the bromoalkaene model, with masses of relatively low information content, shows that there are exceptions. The use of the binary encoded data was necessary to resolve the inherent data structure of the data set, which evidently was concealed in the full intensity spectral data. The binary encoding transformation of spectral data is a powerful manner of achieving separation of classes and is conveniently visualized in the geometry of Hamming spaces.

The successful application of SIMCA statistical pattern recognition on a small microcomputer with limited memory to the small, but realistic, set of organic pollutants shows the power of the technique and the availability of useful information in the low-resolution mass spectra. The applicability of the procedures to actual field data might be questioned since reference spectra of limited numbers of compounds were used in the training sets. However, these particular compounds are in fact those that are being routinely sought in field samples using reverse search spectral matching techniques. Also the 17 masses chosen for the pattern recognition studies contained high information content with respect to a much larger set of compounds than the 78 studied. With the use of gas chromatographic retention time information as an efficient prefilter prior to pattern recognition studies it also may be possible to identify some individual compounds in the set.

ACKNOWLEDGMENT

We wish to thank William J. Dunn III of the College of Pharmacy, University of Illinois at Chicago, for many helpful discussions concerning SIMCA pattern recognition. Lynn Wright of EMSL, U.S. EPA, Research Triangle Park, NC, kindly provided the mass spectral files used in this study.

LITERATURE CITED

(1) Martinsen, D. P. *Appl. Spectrosc.* **1981**, *35*, 255–266.
(2) Rozett, R. W.; Petersen, E. McLaughlin *Anal. Chem.* **1975**, *47*, 1301–1308.
(3) Justice, J. B.; Isenhour, T. L. *Anal. Chem.* **1975**, *47*, 2286–2288.
(4) Rozett, R. W.; Petersen, E. McLaughlin *Anal. Chem.* **1976**, *48*, 817–825.
(5) Ritter, G. L.; Lowry, S. R.; Isenhour, T. L. *Anal. Chem.* **1976**, *48*, 591–595.
(6) Malinowski, E. R. *Anal. Chim. Acta* **1982**, *134*, 129–137.
(7) Williams, S. S.; Lam, R. B.; Isenhour, T. L. *Anal. Chem.* **1983**, *55*, 1117–1121.
(8) Hardy, J.; Jardine, I. In "Advances in Mass Spectrometry"; West, A. R., Ed.; Applied Science Publishers: Barking, England, 1974; Vol. VI, pp 1061–1071.
(9) Heller, S. R.; Chang, C. L.; Chu, K. C. *Anal. Chem.* **1974**, *46*, 951–952.
(10) Ziemer, J. N.; Perone, S. P.; Caprioli, R. M.; Seifert, W. E. *Anal. Chem.* **1979**, *51*, 1732–1738.
(11) Lowry, S. R.; Isenhour, T. L.; Justice, J. B.; McLafferty, F. W.; Dayringer, H. E.; Venkataraghavan, R. *Anal. Chem.* **1977**, *49*, 1720–1722.
(12) McGill, J. R.; Kowalski, B. R. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 52–55.
(13) Justice, J. B.; Isenhour, T. L. *Anal. Chem.* **1974**, *46*, 223–226.
(14) Lam, T. F.; Wilkins, C. L.; Brunner, T. R.; Soltzberg, L. J.; Kaberline, S. L. *Anal. Chem.* **1976**, *48*, 1768–1774.
(15) Wold, S.; Christie, H. J. *Anal. Chim. Acta* **1984**, *165*, 51–59.
(16) Wangen, L. E.; Woodward, W. S.; Isenhour, T. L. *Anal. Chem.* **1971**, *43*, 1605–1614.
(17) Van Marlen, G.; Dijkstra, A.; van't Klooster, H. A. *Anal. Chem.* **1979**, *51*, 420–423.
(18) Van Marlen, G.; Dijkstra, A. *Anal. Chem.* **1976**, *48*, 595–598.
(19) Van Marlen, G.; Dijkstra, A.; van't Klooster, H. A. *Anal. Chim. Acta* **1979**, *112*, 233–243.
(20) Van Marlen, G.; van den Hende, J. H. *Anal. Chim. Acta* **1979**, *112*, 143–150.
(21) Dupuis, P. F.; Dijkstra, A.; van der Maas, J. H. *Z. Anal. Chem.* **1978**, *291*, 27–33.
(22) Dupuis, P. F.; Dijkstra, A. *Z. Anal. Chem.* **1978**, *290*, 357–368.
(23) Bink, J. C. W. G.; van't Klooster, H. A. *Anal. Chim. Acta* **1983**, *150*, 53–59.
(24) Dupuis, P. F.; Cleij, P.; van't Klooster, H. A.; Dijkstra, A. *Anal. Chim. Acta* **1979**, *112*, 83–93.
(25) Heite, F. H.; Dupuis, P. F.; van't Klooster, H. A. *Anal. Chim. Acta* **1978**, *103*, 313–321.
(26) Dromey, R. G. *Anal. Chem.* **1976**, *48*, 1464–1469.
(27) Varmuza, K. *Anal. Chim. Acta* **1980**, *122*, 227–240.
(28) Albano, C.; Dunn, W. J., III; Edlund, U.; Johansson, E.; Norden, B.; Sjostrom, M.; Wold, S. *Anal. Chim. Acta* **1978**, *290*, 429–443.
(29) Wold, S. *Pattern Recognit.* **1976**, *8*, 127–134.
(30) Wold, S.; Albano, C.; Dunn, W. J., III; Edlund, U.; Esbensen, K.; Geladi, P.; Hellberg, S.; Johansson, E.; Lindberg, W.; Sjostrom, M. In "Chemometrics, Mathematics and Statistics in Chemistry"; Kowalski, B. R., Ed.; D. Reidel Publishing Co.: Boston, MA, 1984; pp 17–96.
(31) Wold, S. *Technometrics* **1978**, *20*, 127.
(32) Woodruff, H. B.; Lowry, S. R.; Ritter, G. L.; Isenhour, T. L. *Anal. Chem.* **1975**, *47*, 2027–2030.
(33) Grotch, S. L. *Anal. Chem.* **1970**, *42*, 1214–1222.
(34) Pesyna, G. M.; McLafferty, F. W.; Venkataraghavan, R.; Dayringer, H. E. *Anal. Chem.* **1975**, *47*, 1161–1164.

RECEIVED for review October 7, 1985. Accepted December 9, 1985. Although the research described in this article has been funded by the United States Environmental Protection Agency, it has not been subjected to Agency review. It therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred. The mention of trade names or commercial products does not constitute endorsement or recommendation for use.